GroupK_HM3

L.Pernice, E.Ruoppolo, M.Tallone, A.Valentinis

2022-11-20

Contents

FSDS - Chapter 6

Ex 6.12

For the `UN` data file at the book's website (see Exercise 1.24), construct a multiple regression model predicting Internet using all the other variables. Use the concept of multicollinearity to explain why adjusted R^2 is not dramatically greater than when GDP is the sole predictor. Compare the estimated GDP effect in the bivariate model and the multiple regression model and explain why it is so much weaker in the multiple regression model.

Solution

Ex 6.14

The data set `Crabs2` at the book's website comes from a study of factors that affect sperm traits of male horseshoe crabs. A response variable, SpermTotal, is the log of the total number of sperm in an ejaculate. It has $\bar{y} = 19.3$ and $s = 2.0$. The two explanatory variables used in the `R` output are the horseshoe crab's carapace width (CW, mean $18.6\,cm$, standard deviation $3.0\,cm$), which is a measure of its size, and color (1=dark, 2=medium, 3=light), which is a measure of adult age, darker ones being older.

Solution

Ex 6.30

When the values of y are multiplied by a constant c, from their formulas, show that s_y and $\hat{\beta}_1$ in the bivariate linear model are also then multiplied by c. Thus, show that $r = \hat{\beta}_1(s_x/s_y)$ does not depend on the units of measurement.

Solution

Ex 6.42

You can fit the quadratic equation $E(Y) = \beta_0 + \beta_1 x + \beta_2 x^2$ *by fitting a multiple regression model with* $x_1 = x$ *and* $x_2 = x^2$.

(a) *Simulate 100 independent observations from the model* $Y = 40.0 - 5.0x + 0.5x^2 + \epsilon$, *where* X *has a uniform distribution over* $[0, 10]$ *and* $\epsilon \sim N(0, 1)$. *Plot the data and fit the quadratic model. Report how the fitted equation compares with the true relationship.*

(b) *Find the correlation between* x *and* y *and explain why it is so weak even though the plot shows a strong relationship with a large* R^2 *value for the quadratic model.*

Solution

Ex 6.52

F *statistics have alternate expressions in terms of* R^2 *values.*

(a) *Show that for testing* $H_0 : \beta_1 = \cdots = \beta_p = 0$,

$$F = \frac{(TSS - SSE)/p}{SSE/[n - (p + 1)]} \text{ is equivalently } \frac{R^2/p}{(1 - R^2)/[n - (p + 1)]}$$

Explain why larger values of R^2 *yield larger values of* F.

(b) *Show that for comparing nested linear models,*

$$F = \frac{(SSE_0 - SSE_1)/(p_1 - p_0)}{SSE_1/[n - (p_1 + 1)]} = \frac{(R_1^2 - R_0^2)/(p_1 - p_0)}{(1 - R_1^2)/[n - (p_1 + 1)]}$$

Solution

FSDS - Chapter 7

Ex 7.4

Analogously to the previous exercise, randomly sample 30 X *observations from a uniform in the interval* $(-4, 4)$ *and conditional on* $X = x$, *30 normal observations with* $E(Y) = 3.5x^3 - 20x^2 + 0.5x + 20$ *and* $\sigma = 30$. *Fit polynomial normal GLMs of lower and higher order than that of the true relationship. Which model would you suggest? Repeat the same task for* $E(Y) = 0.5x^3 - 20x^2 + 0.5x + 20$ *(same* σ*) several times. What do you observe? Which model would you suggest now?*

Solution

Ex 7.20

In the `Crabs` *data file introduced in Section 7.4.2, the variable* y *indicates whether a female horseshoe crab has at least one satellite (1= yes, 0= no).*

(a) *Fit a main-effects logistic model using weight and categorical color as explanatory variables. Conduct a significance test for the color effect, and construct a 95% confidence interval for the weight effect.*

(b) *Fit the model that permits interaction between color as a factor and weight in their effects, showing the estimated effect of weight for each color. Test whether this model provides a significantly better fit.*

(c) *Use AIC to determine which models seem most sensible among the models with*

(i) interaction,

(ii) main effects,

(iii) weight as the sole predictor,

(iv) color as the sole predictor,

(v) the null model.

Solution

Ex 7.26

A headline in `The Gainesville Sun (Feb. 17, 2014)` *proclaimed a worrisome spike in shark attacks in the previous two years. The reported total number of shark attacks in Florida per year from 2001 to 2013 were 33, 29, 29, 12, 17, 21, 31, 28, 19, 14, 11, 26, 23. Are these counts consistent with a null Poisson model? Explain, and compare aspects of the Poisson model and negative binomial model fits.*

Solution

DAAG - Chapter 8

Ex 6

Sugar content in cereal is monitored in two ways: a lengthy lab analysis and by using quick, inexpensive high performance liquid chromatography. The data in `frostedflakes` *(DAAG) come from 101 daily samples of measurements taken using the two methods.*

(a) *Obtain a vector of differences between the pairs of measurements.*
(b) *Plot the sample autocorrelation function of the vector of differences. Would an MA(1) model be more realistic than independence?*
(c) *Compute a confidence interval for the mean difference under the independence assumption and under the MA(1) assumption.*

Solution